DOCUMENT RESUME

ED 152 805 95 TM 006 965

AUTHOR	Pugh, Richard C.; And Others
TITLE	Categorizing Teacher Behavior: Test Manual.
INSTITUTION	Indiana Univ., Bloomington. Center for Development in Teacher Education.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	76
NOTE	18p.; For related document, see ED 123 254
AVAILABLE FROM	Audio-Visual Center, Indiana University, Bloomington, Indiana 47401 ($1.00)
EDRS PRICE	MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS	Answer Keys; Answer Sheets; *Classification; *Criterion Referenced Tests; Higher Education; *Instructional Films; *Manuals; *Protocol Materials; Questioning Techniques; Sound Films; Student Teacher Relationship; *Teacher Behavior; Teacher Education; Test Reliability; Test Validity; Visual Measures
IDENTIFIERS	*Categorizing Teacher Behavior

ABSTRACT

Categorizing Teacher Behavior, developed to measure acquisition of concepts portrayed in the Concepts and Patterns in Teacher-Pupil Interaction film series, measures the ability to identify and categorize instances of the specificed concepts as they are portrayed in reasonably complex filmed classroom interactions. The film series itself is based upon three pairs of concepts referring to teacher behavior in an interactive setting. These pairs of concepts are: probing and informing, approving and disapproving, productive and reproductive questioning. The test consists of 30 classroom episodes, each less than one minute in length, selected from "out-takes" from the films in the series (that is, from footage not included in the instructional films themselves. In two studies using graduate educational psychology students, the test evidenced a high degree of reliability and validity. Directional predictions were confirmed using composite test scores and composite half-test scores in most instances. A discernible and significant gain in mean test performance can be accounted for by the instruction through the use of protocol films. An answer key and answer form are appended. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED152805

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Patricia Wetmore
Indiana University

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Competency Based Education Manual

Richard C. Pugh
Gordon C. Wilson
Patricia Wetmore

Center for Development in Teacher Education
Indiana University

Categorizing Teacher Behavior: Test Manual

Richard C. Pugh
Gordon Goodwin
David Gliessman

Developed through the

Center for Development in Teacher Education
Indiana University

Distributed by the

Audio-Visual Center
Indiana University
1976

Bloomington, Indiana

Foreword

This manual accompanies the filmed test, *Categorizing Teacher Behavior*, which is part of the protocol film series, *Concepts and Patterns in Teacher-Pupil Interaction*. The development, production and evaluation of the complete series was conducted by the Indiana University Protocol Materials Project, an adjunct of the National Center for the Development of Training Materials in Teacher Education, through a grant under the Education Professions Development Act as administered by the U.S. Office of Education.

The user has permission to reproduce in any quantity the answer booklet provided in Appendix II. Further information on the availability of the films in the series may be obtained by contacting the Audio-Visual Center, Indiana University, Bloomington, Indiana 47401.

Table of Contents

	Page
General Information	1
Test Rationale	1
Concept Definitions	1
Test Development and Characteristics	1
Test Administration and Scoring	2
Technical Information	3
Definition of Strategy to Assess Test Characteristics	3
Implementation of Strategy	3
Conclusions	5
Response Bias	5
References	6
Appendix I	7
Correct Answers	
Appendix II	8
Answer Booklet	

General Information

This manual provides users of the filmed test, *Categorizing Teacher Behavior*, with sufficient information to use that test in a meaningful manner and to interpret the results of its use accurately. The manual consists of two sections. The present section contains a general description of the test and directions for its administration and scoring. The second section contains evidence on the technical characteristics of the test.

Categorizing Teacher Behavior is an integral part of the protocol film series, *Concepts and Patterns in Teacher-Pupil Interaction.* This film series is based on the premise that the "teacher solicitation or question - pupil response - teacher reaction" sequence is a fundamental one in classroom interaction. The purposes of the series are (1) to specify, define and exemplify a specific set of concepts pertaining to teacher behaviors within this classroom interaction sequence and (2) to provide practice in interpreting classroom behavior in terms of that set of concepts. If the films are used appropriately, they should contribute directly to the acquisition of these concepts.[1]

Test Rationale The test, *Categorizing Teacher Behavior*, was developed to measure acquisition of the set of concepts portrayed in the *Concepts and Patterns* in *Teacher-Pupil Interaction* film series. The principal cognitive outcome that can be measured through the use of the test is the ability to identify and categorize instances of the specified concepts as they are portrayed in reasonably complex filmed classroom interactions. This criterion is consistent with the general purpose of protocol materials: to aid in the functional acquisition of concepts as a basis for interpretive skill.

The development of *Categorizing Teacher Behavior* was central to the evaluation plan for the entire film series. The "logic" of developing an entirely film-based protocol series led quite reasonably to the specification that the principal evaluation instrument also be film-based. It was felt that the task of categorizing behavior portrayed on film would most closely approximate the interpretation of behavior in actual classrooms.

Concept Definitions The film series itself is based upon three pairs of concepts referring to teacher behavior in an interactive setting. These pairs of concepts are: probing and informing, approving and disapproving, productive and reproductive questioning.[2] The filmed test provides the instructor or evaluator with direct evidence on the question: What level of competence does each student have in the functional understanding of these concepts? In other words, what level of competence is achieved by each student in reliably categorizing behavior in terms of these concepts? Following are definitions of the concepts themselves:

PROBING A teacher reaction in the form of a question or implied question that pursues some aspect of the substantive content of a preceding student response. Such probes typically seek a clarification, extension or evaluation of that substantive content. By "preceding response" is meant *any* preceding response including, but not limited to, the student response that immediately precedes a teacher reaction. By "substantive content" is meant the formal content of classroom discussion as opposed to such procedural content as assignment making, the order of discussion, or disciplinary matters.

INFORMING A teacher reaction in which information is introduced or reintroduced that is related to some aspects of the substantive content of a preceding student response. Such a reaction is often intended to produce some modification in the substantive content of that student response. By "preceding response" is meant *any* response that immediately precedes a teacher reaction. By "substantive content" is meant the formal content of classroom discussion as opposed to such procedural content as assignment making, the order of discussion, or disciplinary matters.

APPROVING A verbal and/or nonverbal teacher reaction that is intended to encourage, or might reasonably be expected to encourage, continued student responding or a continuation of student behavior.

DISAPPROVING A verbal and/or nonverbal teacher reaction that is intended to discourage, or might reasonably be expected to discourage, continued student responding or a continuation of student behavior.

PRODUCTIVE QUESTION A teacher question that is intended to encourage the production of ideas or combinations or ideas as opposed to simply the reproduction of specifically learned content. A student response to such a question may reflect the recall of specifically learned content, but that content is used in such processes as interpretation, application, and evaluation.

REPRODUCTIVE QUESTION A teacher question intended to elicit directly the recall of content specifically learned as part of a course or topic of study. In response to such a question, the student is expected to reproduce accurately such content or to recognize when it is accurately reproduced by someone else.

Test Development and Characteristics The test consists of 30 brief classroom episodes each less than one minute in length. These episodes were selected from "out-takes" from the films in the series (that is, from footage *not* included in the *Concept* or *Pattern* films themselves). As a consequence, the classroom settings, subject matter content, teachers and pupils are the same as in the *Concept* and *Pattern* films. However, the *specific episodes* appearing in the test film do *not* appear in the other films of the series.

[1] The rationale for the film series is more fully developed in *Concepts and patterns in teacher-pupil interaction: a protocol film series*, a manual available from the Audio-Visual Center, Indiana University, Bloomington.

[2] The specific order in which the concepts are treated in this manual resulted from the way in which they were "paired" in the answer booklet itself. No significance should be attached to the fact that this order differs from that used in the *Concepts and Patterns* users manual.

Certain facts about the use of film as a medium in test development should be noted by the user. Although all episodes for the test were drawn from scenes recorded on film and although the test is distributed as a film, the filmed footage was transferred to videotape for purposes of test development and preliminary trial. Consequently, a considerable amount of the data reported in this manual is based on a videotaped version of the test. The use of videotaped preliminary versions for the gathering of data is easily explained. As with printed tests, film-based tests must undergo a number of revisions, and as a medium, film is both expensive and difficult to revise. "Videotape transfers," on the other hand, are both economical and convenient to modify. The authors' expectation was that since the changes introduced by this shift in medium were primarily technical (for example, absence or presence of color and relative sharpness of image), the magnitude and pattern of quantitative results would not be substantially different for the two media formats. A comparison of the results obtained for the videotaped and final filmed formats of the test confirms this expectation.

The test is divided into three parts, providing for all possible combinations of the pairs of concepts. Part I requires that episodes be categorized in terms of probing, informing, approving and disapproving. In Part II, the concepts involved are productive questioning, reproductive questioning, approving and disapproving. In Part III, the concepts are probing, informing, productive questioning and reproductive questioning. Ten episodes are included in each part and separate items for each of four concepts are presented for each episode. In the case of each item, a "yes" or "no" response option is provided for the question of whether or not a specified concept is illustrated in the episode. As a consequence, a total of 120 items (20 items for each concept) are contained in the test.

During development, the test episodes were subjected to scrutiny by six members of a panel, each of whom responded to the episodes independently and then deliberated together until unanimous agreement was reached on the correct response to each item. Only items for which unanimous agreement was reached were retained. From the set of items and episodes emerging from this refinement, episodes were randomly selected for each of the major parts of the test.

Each episode is presented twice with a very brief delay between presentations; the purpose of this repeated presentation is to minimize dependence on memory of the specific episode as a factor in the measurement of concept acquisition. Following the second presentation of each episode, a longer delay is provided during which the examinee responds by recording his or her answer on a separate answer sheet. The total time for the test (including repeated episodes, delays between repeated episodes, and delays for responding) is approximately 45 minutes.

The restriction of items for each episode to four rather than six concepts was planned to achieve a balance between "stimulus simplicity" and "stimulus overload." The resulting "response task" was thought to be at an optimum level of difficulty. Overall, the examinee is required to indicate the presence or absence of each concept 20 times. The limitation of the test to a total of 120 items provides for a reasonably short test administration time as well as a reasonable opportunity to demonstrate concept acquisition.

Table 1

Number of Instances of Concepts By Part Of Test

Concept	Part of Test			
	Part I	Part II	Part III	Total
Probing	4	-	4	8
Informing	4	-	3	7
Approving	6	6	-	12
Disapproving	3	6	-	9
Productive Questioning	-	5	5	10
Reproductive Questioning	-	3	3	6
Total	17	20	15	52

The frequency of occurrence of each concept for each part of the test is presented in Table 1. There are 20 episodes within the total test in which each concept might have occurred and 10 episodes within each part of the test in which each concept might have occurred. By comparing actual frequency of occurrence to the number of episodes (in each of which a given concept *might* occur) one can obtain the ratio of instance to non-instance items for each part and concept.

Table 2

Number Of Episodes With A Given Number Of Concepts

Number of Concepts	Number of Episodes			
Per Episode	Part I	Part II	Part III	Total
1	4	3	6	13
2	5	4	3	12
3	1	3	1	5
Total	10	10	10	30

In terms of a potential source of response bias, an ideal ratio of instance to non-instance items might have been 50:50. However, the method of item selection precluded obtaining an ideal ratio. In the case of this test, the representativeness of the items and episodes was considered to be very important even though the ratios of instance to non-instance correct answers for selected parts of the test deviated from the ideal. A "yes" or "no" response on a given item has about the same probability of being correct for "productive questioning" (10 yes to 10 no correct answers), and "disapproving" (9 yes to 11 no correct answers). For "approving" the ratio is 12 yes to 8 no correct answers; for "probing," the ratio is 8 yes to 12 no; for "informing" it is 7 yes to 13 no; for "reproductive questioning" it is 6 yes to 14 no.

The number of concepts instanced per episode is presented in Table 2. The typical number of concepts instanced for each episode is one or two with a few episodes having three concepts instanced.

Test Administration and Scoring In administering the test, the examinee must be provided with an answer booklet (See Appendix II). The directions are stated on the front of the answer booklet; the examiner should make certain that these directions are clearly understood. The importance of answering every item should be stressed. With a yes/no response format, an examinee has a fifty percent chance of guessing the correct answer. Thus, in scoring the test, the instructor should approximate the guessing process for any omitted item by flipping a coin. If all items are answered, this process is unnecessary.

To assist in interpreting the results of the test, some "baseline data" might be useful. In preliminary studies at Indiana University using the videotaped format of the test, students in graduate level courses in educational psychology who were untrained in the concepts typically scored between 90 and 95

out of a maximum correct socre of 120.[3] After film-based training, students typically scored between 101 and 107. The scores from the untrained classes were well above chance level (chance level = 60); it should be understood that in this first study, however, the untrained classes were given the concept definitions prior to taking the test. Scores received after training reflected significant gains over those of untrained students or of the same students before training. In general, there was a clear trend toward mastery of the concepts. An elaboration of performance on the test and complementary results are reported in the next section of this test manual.

The final filmed format of the test was used with one group of graduate level students; the findings from this trial are reported under results of the second study in the technical information section. The pretest mean for this educational psychology class was 83; after film-based training, the mean was 103. In this second study, concept definitions were *not* provided before pretesting. The correct answer to each of the 120 items is reported in Appendix I, Table 9. An answer key is also reported for a machine scored answer sheet based on a response of "a" for yes and "b" for no.

Technical Information

Categorizing Teacher Behavior was designed to provide a basis for evaluating the effectiveness of the *Concepts and Patterns in Teacher-Pupil Interaction* film series. The test was criterion-referenced because the content was derived from the meaning of the six underlying concepts (probing, informing, approving, disapproving, productive questioning and reproductive questioning) and a mastery performance level was sought. Since there are special problems in the appropriateness of some of the conventional test characteristic indices, a strategy developed by Brown and Pugh (1975) was adopted to portray the characteristics of the test.

Definition of Strategy to Assess Test Characteristics

The strategy used to assess the test characteristics assumed that the objective of the film-based instructional treatment was to bring the learner from a pre-treatment state below the criterion performance level to criterion level. It also assumed that if *Categorizing Teacher Behavior* detects this change of learning state consistently over the six concepts in the instructional setting, over variations in the instructional setting, over separate groups of learners, and over half-tests within the test, then one has evidence both for (1) the reliability and validity of the test and (2) the reliability and validity of the instructional treatment.

The steps taken to establish the content validity of the test were the (1) development of substantial numbers of episodes exemplifying the concepts, (2) arrangement for independent judging of the items by a jury of staff members who had acquired the concepts, (3) equalization of the number of items for each concept, and (4) inclusion of a sufficient number of items so that each half-test was meaningful.

Pre- and post-treatment states. In the development of this test, it was most reasonable to assume a pre-treatment state reflecting some amount of prior learning. Because the concepts are related to commonly treated content in professional teacher training, it was plausible to conclude that they may have been partially acquired prior to treatment. For the pre-treatment state, the mean of the test was expected to be above the 50 percent (or chance) level of difficulty; it was also expected that the variance would be greater than at the post-treatment state. The latter state was, consequently, expected to be characterized by a higher mean and attenuated variance.

Predicted relationships between treatment states. Since the test was to be administered to subjects for whom some prior learning may have occurred, certain relationships between treatment states (i.e., pretreatment and post-treatment states) were predicted. More specifically, repeated measurements across the two treatment states should show an increase in means and a decrease in variance. These directional predictions should hold for test composite scores across all concepts and half-test composite scores. All of these composite comparisons should show statistical significance. The directional predictions should also hold for test concept scores and half-test concept scores; however, the actual directional comparisons may not be individually statistically significant using test concept and half-test concept scores.

All of these predictions were directional and assumed a comparison between groups of students who received effective training on the concepts through the use of the *Concepts and Patterns* film series and either (1) the same groups in a pre-treatment state or (2) an independent pre-treatment group. An additional set of predictions was stated involving comparisons between groups *not* receiving training or comparisons between trained groups. In this case, if no consistent differences were found, additional evidence of test reliability and validity would be demonstrated.

Implementation of Strategy *Testing procedure.* As explained earlier in this manual, the test requires the examinee to identify and categorize episodes of teacher behavior. The task for the examinee is to determine whether or not the teacher's behavior in a given episode exemplifies each of a selected four of the six concepts. A different combination of four "paired" concepts is presented in each of the three parts of the test.

Scoring procedure. In the case of the first study reported below, responses were scored as correct (that is, an accurate categorization of the teacher's behavior); partially correct/partially incorrect (that is, an incomplete categorization of the teacher's bheavior); or incorrect (an inaccurate categorization of the teacher's behavior). Two points were allowed for each correct response and one point for each partially correct/partiaily incorrect response. A perfect score was 240 points and a chance-level score was 120 points. A perusal of examinee's answers indicated, however, that the partially correct/partially incorrect response was seldom used. As a result, this option was discontinued for the second study. In the second study, all items were scored simply as correct or incorrect; thus, a perfect score was 120 points.

Instructional treatment. The *Concepts and Patterns* protocol films were used as part of regular classroom instruction in several intact sections of a graduate level course in educational psychology; these sections are referred to as the "trained sections" in the report of results. The students enrolled in these sections were representative of those normally enrolled in the course: heterogeneous in age, experience and ability. The topics of teacher behavior and teacher-pupil interaction were a conventional part of the course as taught by the instructors involved. The use of the term "untrained" refers to the planned absence of use of the protocol films. All students may well have gained some general familiarity with one or more of the concepts in the series from instruction in previous courses. All students in the first study were familiarized also with the specific concept definitions during a five-minute orientation prior to administration of the test. Thus, pretest performance in the first study was potentially influenced by *both* unplanned and planned prior learning; this influence was evidenced by average pretest scores that were well above chance level. In the second study, the pretest orientation to the concept definitions was omitted; as a result, average pretest scores were closer to chance level.

Data analysis. Designs were constructed which allowed for the statistical test of directional as well as nondirectional comparisons using the total composite score and odd/even split-half composite scores. For the directional comparisons, it was predicted that means would consistently increase and variances would consistently decrease from pre-instruction to post-instruction conditions. Further, concept scores would follow a similar directional pattern. The analysis is reported separately for each of the two studies.

First study design and results. The design used for the first study was a variation of the separate-sample pretest-posttest design with random assignment, within course sections, into two groups. This design was appropriate to the field setting and had superior external validity characteristics with the effects of pretesting and interaction of testing with treatment controlled. The design is depicted as follows:

A	X	B
	X	C

From this design, two sets of directional contrasts of means were predicted, $B > A$ and $C > A$. Conversely, the directionality of contrasts of variances were predicted to be $B < A$ and $C < A$. A nondirectional contrast, $B = C$, was also tested.

[3] In the following section on "Technical Information," in the first study reported, individual scores are multipled by two and the results summarized in terms of that scoring system. Thus, in referring back to the score on this page, the user may wish to double each score from this page.

Table 3

Means And Standard Deviations For Composite Scores

Composite	A (Pre)			B (Post)			C (Post only)		
	n	$\bar{X}$	sd	n	$\bar{X}$	sd	n	$\bar{X}$	sd
Odd	41	94.2	11.5	41	104.2	6.5	48	98.7	7.0
Even	41	99.2	10.2	41	108.5	5.9	48	104.3	7.5
Total	41	193.3	20.2	41	212.7	10.1	48	203.0	12.3

Table 4

Descriptive Summary Of Concept Scores

Concept Scores	Number of items	A (Pre) (n = 41)		B (Post) (n = 41)		C (Post only) (n = 48)	
		$\bar{X}$	sd	$\bar{X}$	sd	$\bar{X}$	sd
Total							
Probing	20	30.4	3.6	35.2	2.8	34.3	3.6*
Informing	20	30.5	5.1	34.7	2.4	33.5	3.5
Approving	20	32.3	4.0	33.6	2.7	32.0*	3.5
Disapproving	20	33.5	4.6	36.1	3.0	33.6	3.2
Productive Questioning	20	33.3	6.4	36.3	2.8	34.4	3.1
Reproductive Questioning	20	33.3	6.1	36.8	2.8	35.1	3.3
Odd							
Probing	10	15.6	2.3	18.6	1.8	17.8	2.5*
Informing	10	14.2	3.0	16.3	2.0	15.9	2.2
Approving	10	14.8	2.0	15.4	1.5	14.4*	2.3*
Disapproving	10	16.9	2.5	18.1	1.9	17.3	1.8
Productive Questioning	10	16.9	4.1	18.0	1.9	16.7*	2.2
Reproductive Questioning	10	15.8	4.1	17.8	2.3	16.6	2.4
Even							
Probing	10	14.8	2.0	16.6	1.7	16.5	2.2*
Informing	10	16.3	2.8	18.4	1.7	17.6	2.3
Approving	10	17.5	2.6	18.2	2.1	17.6	2.0
Disapproving	10	16.5	2.6	18.0	1.9	16.3*	2.4
Productive Questioning	10	16.5	3.0	18.3	1.9	17.7	2.2
Reproductive Questioning	10	17.5	2.8	19.0	1.4	18.5	1.8

*Reversal of prediction relative to A

Table 5

Comparisons Of Total Concept Scores And Odd/Even Concept Scores

	Total Concept Scores		Odd/Even Concept Scores	
	n_H	p_H	n_H	p_H
B-A Comparisons				
Means	6	.02	12	.0002
	6	.02	12	.0002
C-A Comparisons				
Means	5	.11	9	.07
sds	5	.11	9	.07

A summary of the descriptive statistics is reported in Table 3.

For the sets of directional contrasts of means and variances, the predicted trends using odd/even split-half composite scores and total composite scores were found. The B < A contrasts, which involved correlated data, showed significant differences in variance, [t (39) = 2.26 to 4.11, $p < .05$]. After using logarithmic transformations to obtain homogeneous variances, the B > A contrasts of mean differences were significant, [F (1, 40) = 29.28 to 51.24, $p < .001$]. The C < A contrasts of differences in variance [F (40, 47) = 1.87 to 2.70, $p < .05$] and C > A contrasts of differences in means [t (64 to 72) = 2.17 to 2.64, $p < .05$] were significant. Since these contrasts involved independent data, t-tests using separate sample estimates of the population variance were used to test mean differences.

The nondirectional contrasts, B = C, resulted in nonsignificant [F (47, 49) = 1.16 to 1.62, $p > .10$] variance differences but mean differences were consistently significant [t (87) = 2.91 to 3.97, $p < .01$] for odd/even split-half composite scores and total composite scores. For each contrast of means, B was greater than C, a difference that was attributed to the pretesting experience of the B Ss.

In addition, concept scores and odd/even split-half concept scores were generated and the directionality of mean and variance differences were compiled. These results are reported in Table 4.

As reported in Tables 4 and 5, all the B - A contrasts of concept scores and odd/even split-half concept scores were in the predicted direction. That is, all variances were lower after instruction and all means were higher after instruction. With six concept scores and twelve split-half concept scores for both means and variances, 36 of 36 *a priori* predicted changes were found. For the C - A contrasts, 10 of 12 concept score predictions and 18 of 24 split-half concept score predictions were supported.

Second study design and results. The final filmed version of the test was used in a graduate level educational psychology course before and following the use of the *Concepts and Patterns* protocol films in classroom instruction. While it is acknowledged that the effects of pretesting and interaction of testing with treatment were *not* controlled, the absence of changes in the means and variances would raise serious questions about the test and/or the effectiveness of the films. For this reason, the results are reported as evidence for the characteristics of the filmed version of the test.

The examinees who took the test as a pretest were simply introduced to the concept names as part of the test; no specific orientation on the concepts was provided, as had been done in the first study. It was expected that the means would increase and the variances would decrease from pretest to posttest. A summary of the descriptive statistics is provided in Table 6.

Table 6

Means And Standard Deviations For Composite Scores

Composite	A (Pre)			B (Post)		
	n	$\bar{x}$	sd	n	$\bar{x}$	sd
Odd	19	42.5	8.2	19	51.5	4.2
Even	19	41.0	11.9	19	51.3	4.8
Total	19	83.5	19.4	19	102.8	8.1

Table 7

Descriptive Summary Of Concept Scores

Concept Scores	Number of items	A (Pre) (n = 19)		B (Post) (n = 19)	
Total		$\bar{x}$	sd	$\bar{x}$	sd
Probing	20	13.3	2.7	17.9	1.9
Informing	20	13.4	3.9	16.7	1.6
Approving	20	14.6	3.3	15.8	2.2
Disapproving	20	15.9	4.2	17.3	1.4
Productive Questioning	20	14.5	3.9	17.9	2.2
Reproductive Questioning	20	11.8	5.5	17.1	3.7
Odd					
Probing	10	6.3	1.5	9.3	1.0
Informing	10	6.2	2.1	7.5	1.0
Approving	10	7.2	1.5	7.6	1.3
Disapproving	10	8.1	2.3	9.1	0.6
Productive Questioning	10	7.5	1.7	8.8	1.2
Reproductive Questioning	10	4.8	3.6	8.2	2.4
Even					
Probing	10	7.0	1.7	8.6	1.2
Informing	10	7.3	2.1	9.2	1.0
Approving	10	7.5	2.3	8.2	1.3
Disapproving	10	7.7	2.1	8.5	1.0
Productive Questioning	10	7.0	2.5	8.9	1.6
Reproductive Questioning	10	7.0	2.3	8.8	1.5

Table 8

Comparisons Of Total Concept Scores And Odd/Even Concept Scores

	Total Concept Scores		Odd/Even Concept Scores	
	n_H	p_H	n_H	p_H
Means	6	.02	12	.0002
sds	6	.02	12	.0002

The predicted trends were found using odd/even split-half composite scores and total composite scores. The correlated variances of the post-test scores were significantly lower than the pretest scores [t (17) = 3.76 to 5.96, $p < .01$]. The means of the posttest scores were significantly higher than the pretest scores [F (1, 18) = 22.58 to 35.17, $p < .001$].

The concept scores and odd/even split-half concept scores of the pretest were compared with the same scores on the posttest. The predicted directionalities of means and variances were found and are reported in Table 7. The probability of obtaining each number of predicted differences by chance alone is reported in Table 8.

From Tables 7 and 8, it can be seen that all the B - A contrasts of total concept scores and odd/even split-half concept scores were in the predicted direction. The means from the posttest were higher than the means from the pretest, and the variances from the posttest were lower than the variances from the pretest. Although several alternative explanations might be offered, the test consistently gave indications of being sensitive to instruction through the use of the *Concept and Pattern* films.

Conclusions The following conclusion regarding the test characteristics of *Categorizing Teacher Behavior* seems warranted: by virtue of the proportion of predictions that were substantiated, the test evidenced a high degree of reliability (consistency) and validity. Directional predictions were confirmed using composite test scores and composite half-test scores in most instances. There clearly was a discernible and significant gain in mean test performance which can be accounted for by instruction through the use of the protocol films. Further positive evidence on the characteristics of the test was found using full test concept scores and half-test concept scores. The overwhelming majority of these comparisons substantianted *a priori* predictions.

Response Bias As a supplementary analysis to the strategy of assessing the technical characteristics of the test, response bias was investigated. The test was constructed with the realization that, even without knowledge of the concepts, examinees could conceivably score higher than chance level by biasing their responses; that is, one could try to "second guess" the test. This possibility is due primarily to the fact that instancing of the concepts was done with a view to approximating a realistic classroom setting rather than constructing an ideal correct response distribution.

It is quite reasonable to conclude that longer classroom episodes might evidence a greater number of concepts. In fact, even though all episodes are less than one minute in length, the longer episodes do tend to illustrate more concepts. There was a zero-order correlation coefficient of .37 ($df = 28$, $p < .05$) between the length of the episode (measured in seconds) and the number of concepts instanced. Since this relationship was evident, the tendency for an examinee who has received

concept training to respond "yes" in the longer episodes was investigated by computing the average number of "yes" responses for each episode in the first study. Further, a first-order partial correlation coefficient was computed between the length of the episode (in seconds) and the average number of "yes" responses per episode with the *correct* "yes" responses partialled out. The first order partial correlation of .11 [$t(27) = .55$, $p > .05$] was insignificant. This finding indicated that a "yes" response bias in the case of longer episodes was unlikely. It was concluded that "score inflation" by such response behavior was not present in these studies.

Another possibility for examinees in biasing responses was to respond to each item with "no." A person who used this response pattern could score somewhat higher than chance level (fifty percent correct) since concepts are more often absent than present. It was felt that the general and often demonstrated tendency to respond "yes" when in doubt made the possibility of this type of bias negligible.

References

Brown, L. D. & Pugh, R. C. A strategy for assessing the reliability of criterion-referenced tests. Paper presented at the *83rd Annual Convention of the American Psychological Association*, 1975.

Concepts and patterns in teacher-pupil interaction: a protocol film series. Bloomington, Indiana: National Center for the Development of Training Materials in Teacher Education and Audio-Visual Center, 1976.

APPENDIX I

Table 9A

CORRECT ANSWERS– HAND-SCORED

Part I			Part II			Part III		
	Item Number	Answer		Item Number	Answer		Item Number	Answer
A	1	Yes	A	41	No	A	81	Yes
	2	No		42	Yes		82	Yes
	3	Yes		43	Yes		82	Yes
	4	No		44	Yes		84	No
B	5	No	B	45	Yes	B	85	No
	6	No		46	No		86	Yes
	7	No		47	No		87	No
	8	Yes		48	No		88	No
C	9	No	C	49	Yes	C	89	No
	10	Yes		50	No		90	No
	11	Yes		51	Yes		91	Yes
	12	No		52	No		92	No
D	13	No	D	53	No	D	93	Yes
	14	No		54	Yes		94	No
	15	Yes		55	No		95	No
	16	No		56	Yes		96	Yes
E	17	Yes	E	57	Yes	E	97	No
	18	Yes		58	No		98	No
	19	No		59	No		99	No
	20	Yes		60	No		100	Yes
F	21	Yes	F	61	No	F	101	No
	22	No		62	No		102	No
	23	Yes		63	Yes		103	Yes
	24	No		64	Yes		104	No
G	25	Yes	G	65	Yes	G	105	Yes
	26	No		66	No		106	No
	27	Yes		67	Yes		107	Yes
	28	No		68	No		108	No
H	29	No	H	69	No	H	109	No
	30	Yes		70	Yes		110	Yes
	31	Yes		71	Yes		111	No
	32	No		72	Yes		112	No
I	33	No	I	73	Yes	I	113	No
	34	No		74	No		114	No
	35	No		75	Yes		115	No
	36	Yes		76	Yes		116	Yes
J	37	No	J	77	No	J	117	Yes
	38	Yes		78	No		118	No
	39	No		79	No		119	Yes
	40	No		80	Yes		120	No

Table 9B

CORRECT ANSWERS– MACHINE-SCORED

Part I			Part II			Part III		
	Item Number	Answer		Item Number	Answer		Item Number	Answer
A	1	(a)	A	41	(b)	A	81	(a)
	2	(b)		42	(a)		82	(a)
	3	(a)		43	(a)		83	(a)
	4	(b)		44	(a)		84	(b)
B	5	(b)	B	45	(a)	B	85	(b)
	6	(b)		46	(b)		86	(a)
	7	(b)		47	(b)		87	(b)
	8	(a)		48	(b)		88	(b)
C	9	(b)	C	49	(a)	C	89	(b)
	10	(a)		50	(b)		90	(b)
	11	(a)		51	(a)		91	(a)
	12	(b)		52	(b)		92	(b)
D	13	(b)	D	53	(b)	D	93	(a)
	14	(b)		54	(a)		94	(b)
	15	(a)		55	(b)		95	(b)
	16	(b)		56	(a)		96	(a)
E	17	(a)	E	57	(a)	E	97	(b)
	18	(a)		58	(b)		98	(b)
	19	(b)		59	(b)		99	(b)
	20	(a)		60	(b)		100	(a)
F	21	(a)	F	61	(b)	F	101	(b)
	22	(b)		62	(b)		102	(b)
	23	(a)		63	(a)		103	(a)
	24	(b)		64	(a)		104	(b)
G	25	(a)	G	65	(a)	G	105	(a)
	26	(b)		66	(b)		106	(b)
	27	(a)		67	(a)		107	(a)
	28	(b)		68	(b)		108	(b)
H	29	(b)	H	69	(b)	H	109	(b)
	30	(a)		70	(a)		110	(a)
	31	(a)		71	(a)		111	(b)
	32	(b)		72	(a)		112	(b)
I	33	(b)	I	73	(a)	I	113	(b)
	34	(b)		74	(b)		114	(b)
	35	(b)		75	(a)		115	(b)
	36	(a)		76	(a)		116	(a)
J	37	(b)	J	77	(b)	J	117	(a)
	38	(a)		78	(b)		118	(b)
	39	(b)		79	(b)		119	(a)
	40	(b)		80	(a)		120	(b)

APPENDIX II

Answer Booklet For

Categorizing Teacher Behavior

Part I

Directions. You are about to see 10 short episodes of teacher-pupil interaction. For each episode, you are to judge whether the teacher *probes, informs, approves* and/or *disapproves*. In fact, *one or more* of these behaviors may occur in a single episode; or *none* of these behaviors may be present. Your job is to identify all (if any) of the teacher behaviors that are present in each episode. Make *no* assumptions regarding what you think occurred *prior* to the beginning of each episode. Attend *only* to the interaction contained within each episode.

General Procedure. Each episode will be shown twice. In the first showing, an alphabetical letter will briefly appear superimposed on the film frame; this letter corresponds to the item on your answer sheet. In the second showing, the letter will not appear. After the second showing, you will have sufficient time to record your answer.

Sequence.

1. You will view a classroom episode with a superimposed alphabetical letter that corresponds to the alphabetically lettered item on your answer sheet.

2. At the end of the episode, a 3 or 4 second pause will occur before you view the same episode for a second time.

3. The episode will appear again.

4. At the end of the second showing, a 15 second pause will occur to give you the opportunity to decide which teacher behaviors you perceived.

5. A new episode will appear with the next consecutive alphabetical letter superimposed; this letter corresponds to the next item on your answer sheet. Procedures 2, 3, and 4 will recur for each subsequent episode. In Part I, there will be 10 episodes labeled "A: through "J."

Responses and Scoring. Decide whether or not you think each behavior that is listed for a given episode *did* or *did not* occur during that episode. Then simply place an "X" beside the YES (did) or the NO (did not) on your answer sheet. You will do this for *four different behaviors* in the case of each episode. Be sure to respond to every item; leave no item blank. Your score will be the number of items which you answer correctly.

Episode A

Item

A1 Probing
_____ Yes
_____ No

A2 Informing
_____ Yes
_____ No

A3 Approving
_____ Yes
_____ No

A4 Disapproving
_____ Yes
_____ No

Episode B

Item

B1 Probing
_____ Yes
_____ No

B2 Informing
_____ Yes
_____ No

B3 Approving
_____ Yes
_____ No

B4 Disapproving
_____ Yes
_____ No

Episode C

Item

C1 Probing
_____ Yes
_____ No

C2 Informing
_____ Yes
_____ No

C3 Approving
_____ Yes
_____ No

C4 Disapproving
_____ Yes
_____ No

Episode D

Item

D1 Probing
_____ Yes
_____ No

D2 Informing
_____ Yes
_____ No

D3 Approving
_____ Yes
_____ No

D4 Disapproving
_____ Yes
_____ No

Episode E

Item

E1 Probing
_____ Yes
_____ No

E2 Informing
_____ Yes
_____ No

E3 Approving
_____ Yes
_____ No

E4 Disapproving
_____ Yes
_____ No

Episode F

Item

F1 Probing
_____ Yes
_____ No

F2 Informing
_____ Yes
_____ No

F3 Approving
_____ Yes
_____ No

F4 Disapproving
_____ Yes
_____ No

Episode G

Item

G1 Probing
_____ Yes
_____ No

G2 Informing
_____ Yes
_____ No

G3 Approving
_____ Yes
_____ No

G4 Disapproving
_____ Yes
_____ No

Episode H

Item

H1 Probing
_____ Yes
_____ No

H2 Informing
_____ Yes
_____ No

H3 Approving
_____ Yes
_____ No

H4 Disapproving
_____ Yes
_____ No

Episode I

Item

I1 Probing
_____ Yes
_____ No

I2 Informing
_____ Yes
_____ No

I3 Approving
_____ Yes
_____ No

I4 Disapproving
_____ Yes
_____ No

Episode J

Item

J1 Probing
_____ Yes
_____ No

J2 Informing
_____ Yes
_____ No

J3 Approving
_____ Yes
_____ No

J4 Disapproving
_____ Yes
_____ No

End Of Part I

Part II

Directions. You are about to see 10 more episodes of teacher-student interaction. For each episode, you are to judge whether the teacher asks a *productive* question, asks a *reproductive* question, *approves* and/or *disapproves*. In fact, one or more of these behaviors may be present. Your job is to identify all (if any) of the teacher behaviors that are present in each episode.

Procedures. The same procedure and sequence as in Part I will be followed.

Responses and Scoring. The same scoring system will be used in Part II as was used in Part I.

Answer Sheet For Part II **Productive Question, Reproductive Question, Approving, Disapproving**

Episode A

Item

A1 Productive Question
_____ Yes
_____ No

A2 Reproductive Question
_____ Yes
_____ No

A3 Approving
_____ Yes
_____ No

A4 Disapproving
_____ Yes
_____ No

Episode B

Item

B1 Productive Question
_____ Yes
_____ No

B2 Reproductive Question
_____ Yes
_____ No

B3 Approving
_____ Yes
_____ No

B4 Disapproving
_____ Yes
_____ No

Episode C

Item

C1 Productive Question
_____ Yes
_____ No

C2 Reproductive Question
_____ Yes
_____ No

C3 Approving
_____ Yes
_____ No

C4 Disapproving
_____ Yes
_____ No

Episode D

Item

D1 Productive Question
_____ Yes
_____ No

D2 Reproductive Question
_____ Yes
_____ No

D3 Approving
_____ Yes
_____ No

D4 Disapproving
_____ Yes
_____ No

Episode E

Item

E1 Productive Question
_____ Yes
_____ No

E2 Reproductive Question
_____ Yes
_____ No

E3 Approving
_____ Yes
_____ No

E4 Disapproving
_____ Yes
_____ No

Episode F

Item

F1 Productive Question
_____ Yes
_____ No

F2 Reproductive Question
_____ Yes
_____ No

F3 Approving
_____ Yes
_____ No

F4 Disapproving
_____ Yes
_____ No

Episode G

Item

G1 Productive Question
_____ Yes
_____ No

G2 Reproductive Question
_____ Yes
_____ No

G3 Approving
_____ Yes
_____ No

G4 Disapproving
_____ Yes
_____ No

Episode H

Item

H1 Productive Question
_____ Yes
_____ No

H2 Reproductive Question
_____ Yes
_____ No

H3 Approving
_____ Yes
_____ No

H4 Disapproving
_____ Yes
_____ No

Episode I

Item

I1 Productive Question
_____ Yes
_____ No

I2 Reproductive Question
_____ Yes
_____ No

I3 Approving
_____ Yes
_____ No

I4 Disapproving
_____ Yes
_____ No

Episode J

Item

J1 Productive Question
_____ Yes
_____ No

J2 Reproductive Question
_____ Yes
_____ No

J3 Approving
_____ Yes
_____ No

J4 Disapproving
_____ Yes
_____ No

End Of Part II

Part III

Directions. You are about to see a final 10 episodes of teacher-student interaction. For each episode, you are to judge whether the teacher *probes, informs,* asks a *productive question* and/or asks a *reproductive question.* Your job is to identify all (if any) of the teacher behaviors that are present in each episode.

Procedures. The same procedure and sequence as in Parts I and II will be followed.

Responses and Scoring. The same scoring system will be used in Part III as was used in Parts I and II.

Answer Sheet For Part III **Probing, Informing, Productive Question, Reproductive Question**

Episode A

Item

A1 Probing
_____ Yes
_____ No

A2 Informing
_____ Yes
_____ No

A3 Productive Question
_____ Yes
_____ No

A4 Reproductive Question
_____ Yes
_____ No

Episode B

Item

B1 Probing
_____ Yes
_____ No

B2 Informing
_____ Yes
_____ No

B3 Productive Question
_____ Yes
_____ No

B4 Reproductive Question
_____ Yes
_____ No

Episode C

Item

C1 Probing
_____ Yes
_____ No

C2 Informing
_____ Yes
_____ No

C3 Productive Question
_____ Yes
_____ No

C4 Reproductive Question
_____ Yes
_____ No

Episode D

Item

D1 Probing
_____ Yes
_____ No

D2 Informing
_____ Yes
_____ No

D3 Productive Question
_____ Yes
_____ No

D4 Reproductive Question
_____ Yes
_____ No

Episode E

Item

E1 Probing
_____ Yes
_____ No

E2 Informing
_____ Yes
_____ No

E3 Productive Question
_____ Yes
_____ No

E4 Reproductive Question
_____ Yes
_____ No

Episode F

Item

F1 Probing
_____ Yes
_____ No

F2 Informing
_____ Yes
_____ No

F3 Productive Question
_____ Yes
_____ No

F4 Reproductive Question
_____ Yes
_____ No

Episode G

Item

G1 Probing
_____ Yes
_____ No

G2 Informing
_____ Yes
_____ No

G3 Productive Question
_____ Yes
_____ No

G4 Reproductive Question
_____ Yes
_____ No

Episode H

Item

H1 Probing
_____ Yes
_____ No

H2 Informing
_____ Yes
_____ No

H3 Productive Question
_____ Yes
_____ No

H4 Reproductive Question
_____ Yes
_____ No

Episode I

Item

I1 Probing
_____ Yes
_____ No

I2 Informing
_____ Yes
_____ No

I3 Productive Question
_____ Yes
_____ No

I4 Reproductive Question
_____ Yes
_____ No

Episode J

Item

J1 Probing
_____ Yes
_____ No

J2 Informing
_____ Yes
_____ No

J3 Productive Question
_____ Yes
_____ No

J4 Reproductive Question
_____ Yes
_____ No

End Of Part III